Infinitely imbalanced binomial regression and deformed exponential families

T. Sei

April 23, 2013

Abstract

The logistic regression model is known to converge to a Poisson point process model if the binary response tends to infinitely imbalanced. In this paper, it is shown that this phenomenon is universal in a wide class of link functions on binomial regression. The proof relies on the extreme value theory. For the logit, probit and complementary log-log link functions, the intensity measure of the point process becomes an exponential family. For some other link functions, deformed exponential families appear. A penalized maximum likelihood estimator for the Poisson point process model is suggested.
Keywords: binomial regression; extreme value theory; imbalanced data; Poisson point process; q-exponential family.

1 Introduction

Let $\{(X_i, Y_i)\}_{i=1}^m$ be m independently and identically distributed observable data on $\mathbb{R}^p \times \{0, 1\}$. The conditional distribution of Y_i given X_i is assumed to be

$$P(Y_i = 1 \mid X_i, a, b) = G(a + b^{\mathrm{T}} X_i), \quad a \in \mathbb{R}, \quad b \in \mathbb{R}^p, \tag{1}$$

where $G(\cdot)$ is a one-dimensional cumulative distribution function. The inverse function $G^{-1}(p) = \sup\{z : G(z) \leq p\}$ is the link function in terms of generalized linear models. Denote the marginal distribution of X_i by $F(dX_i)$. The distribution function G is typically the logistic, standard normal or Gumbel distributions. The corresponding link functions are the logit, probit and

complementary log-log functions, respectively. For the three examples, the log-likelihood function of (1) is concave; see Wedderburn (1976).

Our interest is the situation that the data is highly imbalanced. In other words, the probability of success is almost zero. Examples of such cases are fraud detection, medical diagnosis, political analysis and so forth. See e.g. Bolton & Hand (2002), Chawla et al. (2004), Jin et al. (2005), and King & Zeng (2001). For the data without covariates, Poisson's law of rare events is well known: if $P(Y_i = 1) = \lambda/m + o(m^{-1})$, then the probability distribution of $\sum_{i=1}^m Y_i$ converges to the Poisson distribution with the mean parameter λ. From this observation, for highly imbalanced data, it is natural to consider that the true parameter (a, b) in (1) depends on m, say (a_m, b_m), and $G(a_m) \to 0$ as $m \to \infty$.

Owen (2007) showed that the maximum likelihood estimator of the logistic regression model converges to that of an exponential family if $\sum_{i=1}^m Y_i$ is fixed and m goes to infinity. This result is roughly derived as follows. Consider the model (1) with the logistic distribution $G(z) = e^z/(1 + e^z)$. Take $a_m(\alpha) = -\log m + \alpha$ and $b_m(\beta) = \beta$ for any fixed α and β. Then we obtain

$$P(Y_i = 1 \mid X_i, a_m(\alpha), b_m(\beta)) = \frac{e^{-\log m+\alpha+\beta^{\mathrm{T}}X_i}}{1 + e^{-\log m+\alpha+\beta^{\mathrm{T}}X_i}} = \frac{e^{\alpha+\beta^{\mathrm{T}}X_i}}{m} + o(m^{-1}) \tag{2}$$

as $m \to \infty$. By Bayes' theorem, the conditional density of X_i given $Y_i = 1$ with respect to the distribution $F(dX_i)$ is, at least formally,

$$\frac{e^{\beta^{\mathrm{T}}X_i}}{\int e^{\beta^{\mathrm{T}}x} F(dx)} + o(1). \tag{3}$$

This is an exponential family with the sufficient statistic x_i, and Owen's result follows.

Remark 1. To be precise, Owen (2007) proved the convergence result under a different setting from here. He assumed that the true conditional distribution of X_i given $Y_i = j$, $j \in \{0, 1\}$, is any distribution F_j. In our setting, F_0 is asymptotically equal to F, and the density of F_1 with respect to F should satisfy (3). In other words, our setting becomes misspecified unless this equality is satisfied. We discuss this point again in Section 5.

Warton & Shepherd (2010) pointed out that the likelihood of logistic regression converges to a Poisson point process model with a specific form of

intensity. Indeed, by (2), the probability $P(Y_i = 1, X_i \in A)$ is approximately $m^{-1}\int_A e^{\alpha+\beta^{\mathrm{T}}x}F(dx)$ for any compact subset A of $\mathbb{R}^p$. Therefore, by Poisson's law of rare events, the number of observations X_i for which $X_i \in A$ and $Y_i = 1$ is approximately distributed according to the Poisson distribution with mean $\int_A e^{\alpha+\beta^{\mathrm{T}}x}F(dx)$. This is the Poisson point process with the intensity measure $e^{\alpha+\beta^{\mathrm{T}}x}F(dx)$.

In this paper, we consider the limit of various binomial regression models other than the logistic model. As expected from the result on logistic regression, the limit becomes a Poisson point process. A remarkable fact we prove is that the intensity measure of the point process should be a q-exponential family for some real number q. The q-exponential family, also called the deformed exponential family or α-family, is recently much investigated in the literature of statistical physics and information geometry; see e.g. Amari (1985), Amari & Nagaoka (2000), Amari & Ohara (2011), Naudts (2002), Naudts (2010), and Tsallis (1988). The precise definition is given in Section 2. The proof relies on the theory of extreme values. For example, for the probit or complementary log-log link functions, the limit of binomial regression is the usual exponential family as with the logit link. On the other hand, if G is the Cauchy distribution, then the limit becomes a q-exponential family with $q = 2$. If the uniform distribution is used, $q = 0$.

As a related work, Ding et al. (2011) introduced the t-logistic regression, that uses the q-exponential family for binary response, where $q = t$. In Section 3, we show that the t-logistic regression converges to the q-exponential family if $q \geq 0$.

In Section 4, we study a penalized maximum likelihood estimator on the q-exponential family of intensity measures. For some special cases, the estimator is reduced to a known admissible estimator for the Poisson mean parameter; see Ghosh & Yang (1988).

Some related problems are discussed in Section 5.

2 Imbalanced asymptotics of binomial regression

For each real number q, define the q-exponential function by

$$\exp_q(z) = \begin{cases} e^z, & \text{if } q = 1, \\ [1 + (1-q)z]_+^{1/(1-q)}, & \text{if } q \neq 1, \end{cases} \tag{4}$$

where $[z]_+ = \max(z, 0)$ and $[0]_+^{-1} = \infty$. This is inverse of the Box-Cox transformation. Note that $\exp_q(z) = \infty$ for $z \geq -1/(1-q)$ if $q > 1$. The function $\exp_q(z)$ is convex if and only if $q \geq 0$.

Consider the binomial regression model (1) and put the following assumption on the distribution function G.

Assumption 1. There exist $q > 0$, $c_m \in \mathbb{R}$ and $d_m > 0$ such that

$$G(c_m + d_m z) = \frac{1}{m} \exp_q(z) + o(m^{-1}) \tag{5}$$

as $m \to \infty$ for each $z \in \mathbb{R}$.

In the extreme value theory, it is known that there is no other asymptotic form than (5) as long as it exists; see e.g. de Haan & Ferreira (2006, Theorem 1.1.2 and 1.1.3). The number q controls the lower tail structure of G. For example, the logistic distribution satisfies Assumption 1 with $q = 1$, $c_m = -\log m$ and $d_m = 1$. Other examples including the normal and Cauchy distributions are considered in Section 3.

We define

$$a_m(\alpha) = c_m + d_m \alpha \quad \text{and} \quad b_m(\beta) = d_m \beta \tag{6}$$

for $(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^p$ by using the sequences c_m and d_m that satisfy (5). Denote the probability law of $\{(X_i, Y_i)\}_{i=1}^m$ under the true parameter $(a_m(\alpha), b_m(\beta))$ by $P_{m,\alpha,\beta}$.

Now the asymptotic form like (2) follows from the assumption. Indeed,

$$\begin{aligned} P_{m,\alpha,\beta}(Y_i = 1 \mid X_i) &= G(a_m(\alpha) + b_m(\beta)^{\mathrm{T}} X_i) \\ &= G(c_m + d_m(\alpha + \beta^{\mathrm{T}} X_i)) \\ &= \frac{1}{m} \exp_q(\alpha + \beta^{\mathrm{T}} X_i) + o(m^{-1}). \end{aligned}$$

Therefore, as in the logistic regression, we expect that the binomial regression model with G converges to the Poisson point process under Assumption 1.

We give a lemma before the main result.

Lemma 1. Let $(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^p$. Let A be any compact subset of $\mathbb{R}^p$ such that the function $\exp_q(\alpha + \beta^{\mathrm{T}} x)$ is finite over $x \in A$. Then the following equation holds:

$$P_{m,\alpha,\beta}(Y_i = 1, X_i \in A) = \frac{\lambda(A)}{m} + o(m^{-1}), \tag{7}$$

where $\lambda(A) = \int_A \exp_q(\alpha + \beta^{\mathrm{T}} x) F(dx)$.

The proof of Lemma 1 is given in Appendix.

Theorem 1. Denote the observations X_i for which $Y_i = 1$ by $\{x_i\}_{i=1}^n$. Then, under $P_{m,\alpha,\beta}$, the set $\{x_i\}_{i=1}^n$ converges in law to the Poisson point process with the intensity measure

$$\lambda(dx) = \exp_q(\alpha + \beta^{\mathrm{T}} x) F(dx) \tag{8}$$

as $m \to \infty$. More precisely, we have

$$\lim_{m\to\infty} P_{m,\alpha,\beta}\left(\#\{i \mid x_i \in A_j\} = n_j,\ j = 1, \ldots, J\right) = \prod_{j=1}^{J} \frac{\lambda(A_j)^{n_j} e^{-\lambda(A_j)}}{n_j!} \tag{9}$$

for any positive integer J, non-negative integers n_j and mutually disjoint compact subsets A_j of $\mathbb{R}^p$ such that $\exp_q(\alpha + \beta^{\mathrm{T}} x)$ is finite over $x \in A_j$.

The equation (9) is consistent with the definition of weak convergence of point processes; see Embrechts et al. (1997).

Proof of Theorem 1. Define

$$\begin{aligned} x(A) &= \#\{i \in \{1, \ldots, n\} \mid x_i \in A\} \\ &= \#\{i \in \{1, \ldots, m\} \mid (X_i, Y_i) \in A \times \{1\}\}. \end{aligned}$$

Since $\{(X_i, Y_i)\}_{i=1}^m$ is an independent and identically distributed sequence, the random vector $(x(A_1), \ldots, x(A_J))$ for the disjoint compact subsets $\{A_j\}_{j=1}^J$ is distributed as the multinomial distribution. Then, by Lemma 1 and Poisson's law of rare events, $(x(A_1), \ldots, x(A_J))$ converges to independent Poisson random variables with intensity $(\lambda(A_1), \ldots, \lambda(A_J))$. The proof is completed. □

By Theorem 1, the logistic regression model converges to the Poisson point process model with intensity $\exp(\alpha + \beta^{\mathrm{T}} x) F(dx)$ as Warton & Shepherd (2010) showed.

Definition 1. For each $q \in \mathbb{R}$, we call the set of intensity measures (8) the q-exponential family of intensity measures. Denote the law of the process $\{x_i\}_{i=1}^n$ with respect to (8) by $P_{\alpha,\beta}^{(q)}$.

The q-exponential family of intensity measures is closely related to the q-exponential family of probability measures as follows. Denote the total intensity by

$$\Lambda_q(\alpha, \beta) = \int_{\mathbb{R}^p} \exp_q(\alpha + \beta^{\mathrm{T}} x) F(dx). \tag{10}$$

Assume $\Lambda_q(\alpha, \beta) < \infty$. Then the likelihood of $P^{(q)}_{\alpha,\beta}$ is

$$\frac{e^{-\Lambda_q(\alpha,\beta)}}{n!} \prod_{i=1}^{n} \exp_q(\alpha + \beta^{\mathrm{T}} x_i), \tag{11}$$

where the base measure of n is the counting measure on $\{0, 1, \cdots\}$, and the base measure of x_i for each i is the distribution $F(dx_i)$. In (11), the number n of observed points is marginally distributed according to the Poisson distribution with intensity $\Lambda_q(\alpha, \beta)$. Each point x_i is independently distributed according to the q-exponential family defined by the probability density function

$$\frac{\exp_q(\alpha + \beta^{\mathrm{T}} x_i)}{\Lambda_q(\alpha, \beta)} \tag{12}$$

with respect to $F(dx)$. The q-exponential family is also called the deformed exponential family or the α-family; see Amari & Nagaoka (2000) for the α-family, where $\alpha = 2q - 1$ should be distinguished with the regression coefficient α. It is known that the density (12) is also written as $\exp_q(\theta^{\mathrm{T}} x_i - \psi_q(\theta))$ with appropriate θ and $\psi_q(\theta)$; see e.g. Amari & Ohara (2011). However, we do not use this parametrization since the quantity $\Lambda_q(\alpha, \beta)$ remains in the whole likelihood (11).

We conjecture that the maximum likelihood estimator of the binomial regression model $P_{m,\alpha,\beta}$ converges to that of the Poisson process model $P^{(q)}_{\alpha,\beta}$ under mild conditions. However, we only give experimental results in Section 3. Instead, we study the estimation problem of the limit model $P^{(q)}_{\alpha,\beta}$ in Section 4. See also Section 5 for further discussion.

3 Examples

In this section, we give some examples of distributions G satisfying Assumption 1, and experimental results on the maximum likelihood estimation.

Even if G satisfies Assumption 1, the sequences c_m and d_m are not uniquely determined. A unified choice is known (see Galambos (1987, Theorem 2.1.4–2.1.6)). However, in the following examples, one of possible pairs (c_m, d_m) is explicitly given for each case.

For the logistic distribution and the Gumbel distribution $G(z) = 1 - \exp(-e^z)$ on minimum values, we have

$$q = 1, \quad c_m = -\log m, \quad d_m = 1. \tag{13}$$

For the standard normal distribution, we have

$$q = 1, \quad c_m = -(2\log m)^{1/2} + \frac{\log(\log m) + \log(4\pi)}{2(2\log m)^{1/2}}, \quad d_m = (2\log m)^{-1/2}. \tag{14}$$

See e.g. Galambos (1987, Section 2.3.2). For the Cauchy distribution, we have

$$q = 2, \quad c_m = -m/\pi, \quad d_m = m/\pi. \tag{15}$$

For other examples such as t-distribution and Pareto distributions, refer to Galambos (1987) and Embrechts et al. (1997).

We briefly study the t-logistic regression proposed by Ding et al. (2011). For each real number t, let $G_t(z) = \exp_t(z - \gamma_t(z))$, where $\exp_t$ denotes the q-exponential function with $q = t$ and $\gamma_t(z)$ is uniquely determined by

$$\exp_t(z - \gamma_t(z)) + \exp_t(-\gamma_t(z)) = 1. \tag{16}$$

We call $G_t(z)$ the t-logistic distribution. Uniqueness of $\gamma_t(z)$ follows from strictly monotone property of the q-exponential function. The distribution $G_t(z)$ is symmetric in the sense that $G_t(-z) = 1 - G_t(z)$ since $\gamma_t(-z) = -z + \gamma_t(z)$ by (16). We obtain the following theorem. The proof is given in Appendix.

Theorem 2. The t-logistic distribution G_t satisfies Assumption 1 with $q = \max(t, 0)$.

Table 1 and Table 2 show the experimental results. The sample is

$$(X_i, Y_i) = \begin{cases} (0.4 + 0.4(i-1)/(n-1), 1) & \text{if} \quad i \in \{1, \ldots, n\}, \\ ((i-n-1)/(m-n-1), 0) & \text{if} \quad i \in \{n+1, \ldots, m\} \end{cases} \tag{17}$$

for $n = 10$ and various m's. For the binomial regression models, the estimated regression coefficient $(\hat{a}, \hat{b})$ is normalized by (6). From Table 1, the convergence rate for the probit link is very slow, or may not converge. For the others, the rate is satisfactory.

Table 1: Comparison of the maximum likelihood estimate of the Poisson point process model with $q = 1$ and the binomial regression models. The logit, probit and clogog (complementary log-log) link functions are used. The sample is (17) and n is fixed to 10. The normalizing sequence (c_m, d_m) is (13) and (14).

	Poisson process		logit		probit		cloglog	
m	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
10^2	1.6504	1.1737	1.6883	1.3067	2.0282	2.3030	1.6975	1.1883
10^3	1.6277	1.2246	1.6314	1.2373	1.9070	1.8777	1.6322	1.2260
10^4	1.6256	1.2294	1.6260	1.2307	1.8634	1.6725	1.6260	1.2295
10^5	1.6254	1.2299	1.6254	1.2300	1.8330	1.5642	1.6254	1.2299

4 Estimation of the q-exponential family of intensity measures

We deal with estimation problem of the q-exponential family of intensity measures (8). The maximum likelihood estimator is likely to fail to exist for small sample size n. We propose a penalized maximum likelihood estimator.

We put the following assumption for simplicity.

Assumption 2. The covariate distribution $F(dx)$ is known. The support of F, denoted by $S(F)$, is finite, and is not included in any hyperplane in $\mathbb{R}^p$. The observable data $\{x_i\}_{i=1}^n$ belongs to $S(F)$.

In practice, $F(dx)$ may be replaced with the empirical, or estimated, distribution based on the covariate sample $\{X_i\}_{i=1}^m$ of the original regression problem.

The parameter space is

$$\Theta = \{(\alpha, \beta) \mid 1 + (1-q)(\alpha + \beta^{\mathrm{T}} x) > 0 \text{ for any } x \in S(F)\}. \tag{18}$$

The set Θ is convex and unbounded since it is intersection of half spaces including the set $\{(\alpha, 0) \mid 1 + (1-q)\alpha > 0\}$. Furthermore, Θ is open since $S(F)$ is compact. In terms of convex analysis, Θ corresponds to the polar set of $S(F)$. See Barvinok (2002).

Table 2: Comparison of the maximum likelihood estimate of the Poisson point process model with $q = 2$ and the binomial regression model with the cauchit (inverse of Cauchy) link function. The sample is (17) and n is fixed to 10. The normalizing sequence (c_m, d_m) is (15).

	Poisson process		cauchit	
m	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
10^2	0.8662	0.0667	0.8632	0.0656
10^3	0.8626	0.0673	0.8623	0.0677
10^4	0.8622	0.0680	0.8622	0.0679
10^5	0.8621	0.0680	0.8622	0.0679

We consider a penalized log-likelihood function

$$-\Lambda_q(\alpha, \beta) + \sum_{i=1}^{n} \log \exp_q(\alpha + \beta^{\mathrm{T}} x_i) + \kappa \int \log \exp_q(\alpha + \beta^{\mathrm{T}} x) F(dx), \quad (19)$$

where κ is a non-negative regularization parameter. If $\kappa = 0$, (19) is the log-likelihood function; see (11). The penalty term represents a pseudo-data of size κ distributed according to F. The function (19) is concave with respect to (α, β) if $0 \leq q \leq 1$. Indeed, we can directly confirm that $-\exp_q(z)$ is concave if $q \geq 0$, and that $\log(\exp_q(z))$ is concave if $q \leq 1$.

Definition 2. We call the maximizer of (19) the additive-smoothing estimator.

This estimator has a desirable property as shown in the following example, even if $q = 1$.

Example 1. Let F be a two-point distribution on $\mathbb{R}$ defined by

$$F(x = 0) = p_0 \quad \text{and} \quad F(x = 1) = p_1,$$

where $p_0, p_1 > 0$ and $p_0 + p_1 = 1$. Denote the intensity at $x = 0$ and $x = 1$ by $\lambda_0 = p_0 \exp_q(\alpha)$ and $\lambda_1 = p_1 \exp_q(\alpha + \beta)$, respectively. It is not difficult to show that $(\alpha, \beta) \in \Theta$ corresponds one-to-one with $(\lambda_0, \lambda_1) \in \mathbb{R}_+^2$, where $\mathbb{R}_+$ is the set of positive numbers. Hence the model is equivalent to the

independent Poisson observable model with intensity (λ_0, λ_1), regardless of q. Then the penalized log-likelihood (19) becomes

$$-\lambda_0 - \lambda_1 + n_0 \log \lambda_0 + n_1 \log \lambda_1 + \kappa \left(p_0 \log \frac{\lambda_0}{p_0} + p_1 \log \frac{\lambda_1}{p_1} \right),$$

where n_j denotes the number of observations $x_i = j$, $j \in \{0, 1\}$. The additive-smoothing estimator is $\hat{\lambda}_j = n_j + \kappa p_j$, $j \in \{0, 1\}$. If $\kappa > 0$, then $(\hat{\lambda}_0, \hat{\lambda}_1) \in \mathbb{R}^2_+$ and the estimator $(\hat{\alpha}, \hat{\beta})$ always exists. Furthermore, if $0 < \kappa \leq 1$, this estimator is known to be admissible with respect to the Kullback-Leibler loss function; see Ghosh & Yang (1988, Theorem 1). For the same reason, if $S(F)$ has only $p + 1$ points in $\mathbb{R}^p$, then the additive-smoothing estimator is admissible as long as $0 < \kappa \leq 1$.

Let $q = 1$ and F be any distribution satisfying Assumption 2. Then, since the model (11) is an exponential family, the pair $(n, \bar{x}_n)$ is a sufficient statistic, where $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$ is the sample mean. Indeed, the additive-smoothing estimator should satisfy

$$\Lambda_1(\hat{\alpha}, \hat{\beta}) = n + \kappa \quad \text{and} \quad \frac{\int x e^{\hat{\beta}^{\mathrm{T}} x} F(dx)}{\int e^{\hat{\beta}^{\mathrm{T}} x} F(dx)} = \frac{n \bar{x}_n + \kappa \int x F(dx)}{n + \kappa}. \tag{20}$$

For the maximum likelihood estimator, meaning $\kappa = 0$, the second equation of (20) is consistent with the result of Owen (2007). From the theory of exponential families, the solution to (20) always exists if $\kappa > 0$ since $\int x F(dx)$ belongs to the interior of the convex hull of $S(F)$; see Barndorff-Nielsen (1978, Corollary 9.6). On the other hand, the maximum likelihood estimator fails to exist if $\bar{x}_n$ is a boundary point.

For $q \neq 1$, we provide a similar result on existence. First consider the following example. The pair $(n, \bar{x}_n)$ is not a sufficient statistic any more.

Example 2. Let $q = 0$ and F be a three-point distribution on $\mathbb{R}$ defined by $F(x = j) = 1/3$ for $j \in \{0, 1, 2\}$. Denote the number of observations $x_i = j$ by n_j. We use $\theta = 1 + \alpha$ and $\phi = 1 + \alpha + 2\beta$ as a new parameter. Then the parameter space is $\theta > 0$ and $\phi > 0$. The penalized log-likelihood is

$$-\frac{\theta + \phi}{2} + n_0^* \log \theta + n_1^* \log \frac{\theta + \phi}{2} + n_2^* \log \phi, \tag{21}$$

where $n_j^* = n_j + \kappa/3$. The maximizer $(\hat{\theta}, \hat{\phi})$ of (21) is

$$\hat{\theta} = \frac{2n_0^*(n_0^* + n_1^* + n_2^*)}{n_0^* + n_2^*} \quad \text{and} \quad \hat{\phi} = \frac{2n_2^*(n_0^* + n_1^* + n_2^*)}{n_0^* + n_2^*}.$$

This always belongs to the parameter space if $\kappa > 0$. On the other hand, the maximum likelihood estimator fails to exist if $n_0 = 0$ or $n_2 = 0$.

In general, the following theorem holds. The proof is given in Appendix.

Theorem 3. Let q be any real number and $\kappa > 0$. If Assumption 2 is satisfied, then the additive-smoothing estimator exists almost surely. It is unique if $0 \leq q \leq 1$.

5 Discussion

5.1 Multinomial regression

We studied so far the binomial regression. There are variants of multinomial regression models. The multinomial t-logistic regression proposed by Ding et al. (2011) can be proved to have a limit under imbalanced asymptotics in the same manner as Theorem 2. The author was not aware of more general results. The problem is postponed as a future work.

5.2 Convergence of estimator

We did not study convergence properties of estimators such as the maximum likelihood estimator. Instead we considered the additive-smoothing estimator for the q-exponential family of intensity measures in Section 4.

Owen (2007) showed that the maximum likelihood estimator of the logistic regression converges to that of the exponential family under imbalanced asymptotics. Then a natural conjecture is that the maximum likelihood estimator of the binomial regression model, which is the maximizer of

$$\sum_{i=1}^{m} \left[Y_i \log G(a + b^{\mathrm{T}} X_i) + (1 - Y_i) \log\{1 - G(a + b^{\mathrm{T}} X_i)\} \right],$$

converges to that of the q-exponential family. Note that estimation of (a, b) is equivalent to that of (α, β) via the formula (6). It will be also meaningful

to study convergence of statistical experiments; see van der Vaart (1998) for the terminology.

An estimator corresponding to the additive-smoothing estimator of Definition 2 is the maximizer of

$$\sum_{i=1}^{m} [Y_i \log G(a + b^{\mathrm{T}} X_i) + (1 - Y_i) \log\{1 - G(a + b^{\mathrm{T}} X_i)\}] + \frac{\kappa}{m} \sum_{i=1}^{m} \log\{m G(a + b^{\mathrm{T}} X_i)\}$$

since the additional term converges to $\kappa \int \log \exp_q(\alpha + \beta^{\mathrm{T}} x) F(dx)$ after normalization (6). The estimator is expected to converge as well.

5.3 Misspecified case

We studied asymptotic properties of the binomial regression model under an assumption that the model (1) is true. On the other hand, Owen (2007) put a different assumption, in that the true conditional distribution of the covariate X_i given $Y_i = j$, $j \in \{0, 1\}$, is fixed to some distribution F_j. In this assumption, our setting is asymptotically described as $F_0(dx) = F(dx)$ and $F_1(dx) = \{\exp_q(\alpha + \beta^{\mathrm{T}} x)/\Lambda_q(\alpha, \beta)\} F(dx)$ by (11). In other words, if the true distributions F_j do not satisfy this relation, the model is misspecified.

It is important to consider robustness of estimators under the misspecified assumption. The problem is not so serious if the support of F_1 is included in that of F, since then F_1 is absolutely continuous with respect to the estimated intensity measure $\exp(\hat{\alpha} + \hat{\beta}^{\mathrm{T}} x) F(dx)$, whenever $(\hat{\alpha}, \hat{\beta})$ belongs to the parameter space (18). Otherwise, however, F_1 is not absolutely continuous. In other words, the estimated intensity measure does not allow that the future data x_{n+1} falls into a region. In particular, if the support of F_1 is not assumed a priori, there is risk of such a contradiction.

One may consider to take a distribution F with the full support $\mathbb{R}^p$ in order to contain the support of F_1. However, if $q \neq 1$, we cannot assume such a distribution F since the parameter space (18) becomes $\{(\alpha, 0) \mid 1 + (1 - q)\alpha > 0\}$.

A solution to this problem will be to use a parametric family of F together with a Bayesian prior distribution. For example, let $F(dx) = F(dx \mid \theta)$ be the uniform distribution on the hypercube $[-\theta, \theta]^p$, and assume a prior density on $\theta > 0$. As long as the true $F_1(dx)$ has compact support, we have a chance to detect it since there is a sufficiently large θ such that the support of F_1 is included in that of $F(\cdot \mid \theta)$.

5.4 Bayesian prediction

In the preceding subsection, we considered the Bayesian approach for treating misspecified case. Even if the model is correctly specified, the approach will be fruitful.

In Section 4, we considered the additive-smoothing estimator of (α, β). This is considered as a maximum-a-posteriori estimator if the prior density

$$\pi(\alpha, \beta) = \exp\left(\kappa \int \log \exp_q(\alpha + \beta^{\mathrm{T}} x) F(dx)\right)$$

is adopted. Then additive-smoothing Bayesian prediction can be also defined by the same prior.

In Example 1, we noted that, for special cases of F and κ, the additive-smoothing estimator becomes an admissible estimator with respect to the Kullback-Leibler divergence, shown by Ghosh & Yang (1988). For prediction problem, a class of admissible predictive densities is investigated by Komaki (2004). Together with the additive-smoothing estimator, decision-theoretic properties of the additive-smoothing prediction are of interest.

Acknowledgement

The author thanks to Saki Saito for helpful discussions in the exploratory stage.

A Appendix

A.1 Proof of Lemma 1

Denote the induced probability distribution of $t = \alpha + \beta^{\mathrm{T}} X_i$ by $F^*(dt)$. Let A^* be $A^* = \{\alpha + \beta^{\mathrm{T}} x \mid x \in A\}$. Then A^* is compact since A is. We have

$$\begin{aligned} P_{m,\alpha,\beta}(Y_i = 1, X_i \in A) &= \int_A G(a_m(\alpha) + b_m(\beta)^{\mathrm{T}} x) F(dx) \\ &= \int_A G(c_m + d_m(\alpha + \beta^{\mathrm{T}} x)) F(dx) \\ &= \int_{A^*} G(c_m + d_m t) F^*(dt). \end{aligned}$$

To prove (7), it is enough to show that

$$\int_{A^*} G(c_m + d_m t) F^*(dt) = \frac{1}{m} \int_{A^*} \exp_q(t) F^*(dt) + o(m^{-1}).$$

By Assumption 1, we know $mG(c_m + d_m t) = \exp_q(t) + o(1)$ for each $t \in A^*$. Hence it is enough to show that $mG(c_m + d_m t)$ converges to $\exp_q(t)$ uniformly in $t \in A^*$. However, since $mG(c_m + d_m t)$ is monotone in t and $\exp_q(t)$ is continuous in $t \in A^*$, uniform convergence follows from the general argument; see e.g. Galambos (1987, Lemma 2.10.1)).

Proof of Theorem 2

For each real number q, denote the set of distributions that satisfy Assumption 1 by $\mathcal{D}_q$.

For $t = 1$, elementary calculation shows that $G_1(z) = e^z/(1 + e^z)$. This is the logistic distribution and belongs to $\mathcal{D}_1$.

For $t = 0$, we have $G_0(z) = (1 + z)/2$, $-1 < z < 1$. This is the uniform distribution on $[-1, 1]$ and belongs $\mathcal{D}_0$.

Let $t > 1$. It suffices to show that

$$G_t(z) = [(1 - t)z]^{1/(1-t)} + o((-z)^{1/(1-t)}), \quad z \to -\infty.$$

Indeed, by the condition (16), if $z \to -\infty$, then $\gamma_t(z) \to 0$. Thus

$$\begin{aligned} \exp_t(z - \gamma_t(z)) &= \exp_t(z + o(z)) \\ &= [1 + (1 - t)(z + o(z))]^{1/(1-t)} \\ &= [(1 - t)z]^{1/(1-t)} + o((-z)^{1/(1-t)}) \end{aligned}$$

Hence G_t belongs to $\mathcal{D}_t$.

For $t < 1$, we first show that the support of G_t has the infimum $z_* = -1/(1 - t)$ and that $\gamma_t(z)$ tends to 0 as $z \to z_* + 0$. Note that the t-exponential function $\exp_t(z)$ is continuous in $z \in \mathbb{R}$, strictly increasing over $z > z_*$, and remains 0 over $z \le z_*$. Since $\exp_t(z) > 1$ for any $z > 0$, it must be $\gamma_t(z) \ge 0$ for any $z \in \mathbb{R}$ by (16). Then $\exp_t(z - \gamma_t(z)) > 0$ only if $z > z_*$. Conversely, if $z > z_*$, it must be $\exp_t(z - \gamma_t(z)) > 0$. Indeed, if $\exp_t(z - \gamma_t(z)) = 0$, then $\gamma_t(z) = 0$ by (16), but this contradicts $z > z_*$. To prove $\gamma_t(z) \to 0$ as $z \to z_* + 0$, due to (16), it is sufficient to show that $\exp_t(z - \gamma_t(z)) \to 0$ as $z \to z_* + 0$. This is shown as

$$0 \le \exp_t(z - \gamma_t(z)) \le \exp_t(z) \to 0, \quad z \to z_* + 0.$$

Let $0 < t < 1$ and $z_* = -1/(1-t)$. It suffices to show that

$$G_t(z) = [(1-t)(z-z_*)]^{1/(1-t)} + o((z-z_*)^{1/(1-t)}), \quad z \to z_* + 0. \tag{22}$$

By the definition of z_*, we have

$$\begin{aligned} \exp_t(z - \gamma_t(z)) &= [1 + (1-t)(z-\gamma_t(z))]^{1/(1-t)} \\ &= [(1-t)(z - z_* - \gamma_t(z))]^{1/(1-t)}. \end{aligned} \tag{23}$$

On the other hand, since $\gamma_t(z) \to 0$ as $z \to z_* + 0$, we obtain

$$\exp_t(-\gamma_t(z)) = 1 - \gamma_t(z) + o(\gamma_t(z)). \tag{24}$$

By substituting the two equations to (16), we obtain $\gamma_t(z) = O((z-z_*)^{1/(1-t)}) = o(z - z_*)$. Then (23) implies (22). Hence G_t belongs to $\mathcal{D}_t$.

Finally, let $t < 0$ and $z_* = -1/(1-t)$. We show that G_t belongs to $\mathcal{D}_0$, not $\mathcal{D}_t$. It suffices to show that

$$G_t(z) = (z - z_*) + o(z - z_*), \quad z \to z_* + 0. \tag{25}$$

For the same reason as the case $0 < t < 1$, we have the two equations (23) and (24). By substituting them to (16), we obtain

$$\gamma_t(z) = (z - z_*) - \frac{(z-z_*)^{1-t}}{1-t} + o((z-z_*)^{1-t}).$$

Then (23) implies (25). Hence G_t belongs to $\mathcal{D}_0$.

Proof of Theorem 3

Uniqueness follows from concavity of (19) for $0 \le q \le 1$. We prove the existence result. Since the case $q = 1$ is proved in (20), we assume $q \neq 1$.

In the following, we prove the theorem only for the case that $n = 0$, that is, no data is observed. The case $n \ge 1$ is similarly proved if one notes that $\{x_i\}_{i=1}^n$ is contained in the convex hull of the support of F.

Let F be a discrete distribution with support $\{\xi_j\}_{j=1}^J \subset \mathbb{R}^p$ and put $p_j = F(x = \xi_j) > 0$, $j \in \{1, \dots, J\}$. By assumption, $\{\xi_j\}_{j=1}^J$ is not included in any hyperplane of $\mathbb{R}^p$. The parameter space (18) is written as

$$\Theta = \{(\alpha, \beta) \mid 1 + (1-q)(\alpha + \beta^{\mathrm{T}}\xi_j) > 0,\ j \in \{1, \dots, J\}\}.$$

Note that Θ is an open convex set and the origin $(\alpha, \beta) = (0, 0)$ always belongs to Θ. The penalized log-likelihood is, since $n = 0$,

$$L(\alpha, \beta) = \sum_{j=1}^{J} p_j \left\{ -\exp_q(\alpha + \beta^{\mathrm{T}} \xi_j) + \log \exp_q(\alpha + \beta^{\mathrm{T}} \xi_j) \right\}. \tag{26}$$

By continuity of $L(\alpha, \beta)$ over Θ, it is sufficient to show that $L(\alpha, \beta) \to -\infty$ if (α, β) tends to a boundary point of Θ or (α, β) diverges. Note that if (α_0, β_0) is a boundary point of Θ, then $(t\alpha_0, t\beta_0)$ belongs to Θ for any $0 \leq t < 1$ since the origin does.

We prove the claim for $q < 1$ first, and then $q > 1$.

Let $q < 1$. Fix any boundary point (α_0, β_0) of Θ. Then there is at least one ξ_j such that $\exp_q(\alpha_0 + \beta_0^{\mathrm{T}} \xi_j) = 0$. For such ξ_j's, $\exp_q(t(\alpha_0 + \beta_0^{\mathrm{T}} \xi_j)) \to +0$ as $t \to 1 - 0$. For the other ξ_j's, $\exp_q(t(\alpha_0 + \beta_0^{\mathrm{T}} \xi_j))$ is bounded as $t \to 1 - 0$. Then, by (26), the function $L(t\alpha_0, t\beta_0)$ tends to $-\infty$ as $t \to 1 - 0$.

Let $q < 1$ and fix any $(\alpha_1, \beta_1) \in \Theta \setminus \{(0, 0)\}$ such that $(t\alpha_1, t\beta_1) \in \Theta$ for any $t > 0$. Then it is necessary that $\alpha_1 + \beta_1^{\mathrm{T}} \xi_j \geq 0$ for all j. Since $\{\xi_j\}$ is not contained in a hyperplane, there is at least one ξ_j such that $\alpha_1 + \beta_1^{\mathrm{T}} \xi_j > 0$. For such ξ_j's, we have $\exp_q(t\alpha_1 + t\beta_1^{\mathrm{T}} \xi_j) \to \infty$ as $t \to \infty$. For the other ξ_j's, $\exp_q(t\alpha_1 + t\beta_1^{\mathrm{T}} \xi_j) = \exp_q(0) = 1$. Therefore, by (26), the function $L(t\alpha_1, t\beta_1)$ tends to $-\infty$ as $t \to \infty$, and the case $q < 1$ was completed.

Let $q > 1$. Fix any boundary point (α_0, β_0) of Θ. Then there is at least one ξ_j such that $\exp_q(\alpha_0 + \beta_0^{\mathrm{T}} \xi_j) = \infty$. For such ξ_j's, $\exp_q(t(\alpha_0 + \beta_0^{\mathrm{T}} \xi_j)) \to \infty$ as $t \to 1 - 0$. For the other ξ_j's, $\exp_q(t(\alpha_0 + \beta_0^{\mathrm{T}} \xi_j))$ is bounded as $t \to 1 - 0$. Then, by (26), the function $L(t\alpha_0, t\beta_0)$ tends to $-\infty$ as $t \to 1 - 0$.

Finally, let $q > 1$ and fix any $(\alpha_1, \beta_1) \in \Theta \setminus \{(0, 0)\}$ such that $(t\alpha_1, t\beta_1) \in \Theta$ for any $t > 0$. Then it is necessary that $\alpha_1 + \beta_1^{\mathrm{T}} \xi_j \leq 0$. Since $\{\xi_j\}$ is not contained in a hyperplane, there is at least one ξ_j such that $\alpha_1 + \beta_1^{\mathrm{T}} \xi_j < 0$. For such ξ_j's, $\exp_q(t\alpha_1 + t\beta_1^{\mathrm{T}} \xi_j) \to +0$ as $t \to \infty$. For the other ξ_j's, $\exp_q(t\alpha_1 + t\beta_1^{\mathrm{T}} \xi_j) = \exp_q(0) = 1$. Therefore, by (26), the function $L(t\alpha_1, t\beta_1)$ tends to $-\infty$ as $t \to \infty$, and the case $q > 1$ was completed.

References

Amari, S. (1985). *Differential-geometrical methods in statistics*. Berlin: Springer.

Amari, S. & Nagaoka, H. (2000). *Methods of information geometry (Translations of Mathematical Monographs)*. Oxford University Press.

Amari, S. & Ohara, A. (2011). Geometry of q-exponential family of probability distributions. *Entropy* **13**, 1170–1185.

Barndorff-Nielsen, O. (1978). *Information and exponential families*. Chichester: John Wiley & Sons.

Barvinok, A. (2002). *A course in convexity*. American Mathematical Society.

Bolton, R. J. & Hand, D. J. (2002). Statistical fraud detection: a review. *Statist. Sci.* **17**, 235–249.

Chawla, N. V., Japkowicz, N. & Koltz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter* **6**, 1–6.

de Haan, L. & Ferreira, A. (2006). *Extreme value theory, an introduction*. New York: Springer.

Ding, N., Vishwanathan, S. V. N., Warmuth, M. & Denchev, V. (2011). t-logistic regression. *J. Mach. Learn. Res.* **12**, 1–55.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling extremal events*. Berlin: Springer.

Galambos, J. (1987). *The asymptotic theory of extreme order statistics*. Malarbar: Robert E. Krieger Publishing Company.

Ghosh, M. & Yang, M.-C. (1988). Simultaneous estimation of Poisson means under entropy loss. *Ann. Statist.* **16**, 278–291.

Jin, Y., Rejesus, R. M. & Little, B. B. (2005). Binary choice models for rare events data: a crop insurance fraud application. *Applied Economics* **37**, 841–848.

King, G. & Zeng, L. (2001). Logistic regression in rare events data. *Political Analisis* **9**, 137–163.

Komaki, F. (2004). Prediction of independent Poisson observables. *Ann. Statist.* **32**, 1744–1769.

Naudts, J. (2002). Deformed exponentials and logarithms in generalized thermostatistics. *Physica A* **316**, 323–334.

Naudts, J. (2010). The q-exponential family in statistical physics. *J. Phys.: Conf. Ser.* **201**, 012003.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *J. Mach. Learn. Res.* **8**, 761–773.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Phys.* **52**, 479–487.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Warton, D. I. & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence only data in ecology. *Ann. Applied Statist.* **4**, 1383–1402.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.